www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **45**, No. 3, pp. 419–431 (2020)
DOI: 10.24425/aoa.2020.134058

Research Paper

A Study on the Impact of Lombard Effect on Recognition of Hindi Syllabic Units Using CNN Based Multimodal ASR Systems

Sadasivam UMA MAHESWARI[(1)]*, A. SHAHINA[(1)]
Ramesh RISHICKESH[(1)], A. NAYEEMULLA KHAN[(2)]

[(1)] *Department of Information Technology*
SSN College of Engineering
Chennai, India
*Corresponding Author e-mail: umamaheswaris@ssn.edu.in

[(2)] *School of Computing Science and Engineering*
VIT University
Chennai, India

(*received July 10, 2019; accepted April 26, 2020*)

Research work on the design of robust multimodal speech recognition systems making use of acoustic and visual cues, extracted using the relatively noise robust alternate speech sensors is gaining interest in recent times among the speech processing research fraternity. The primary objective of this work is to study the exclusive influence of Lombard effect on the automatic recognition of the confusable syllabic consonant-vowel units of Hindi language, as a step towards building robust multimodal ASR systems in adverse environments in the context of Indian languages which are syllabic in nature. The dataset for this work comprises the confusable 145 consonant-vowel (CV) syllabic units of Hindi language recorded simultaneously using three modalities that capture the acoustic and visual speech cues, namely normal acoustic microphone (NM), throat microphone (TM) and a camera that captures the associated lip movements. The Lombard effect is induced by feeding crowd noise into the speaker's headphone while recording. Convolutional Neural Network (CNN) models are built to categorise the CV units based on their place of articulation (POA), manner of articulation (MOA), and vowels (under clean and Lombard conditions). For validation purpose, corresponding Hidden Markov Models (HMM) are also built and tested. Unimodal Automatic Speech Recognition (ASR) systems built using each of the three speech cues from Lombard speech show a loss in recognition of MOA and vowels while POA gets a boost in all the systems due to Lombard effect. Combining the three complimentary speech cues to build bimodal and trimodal ASR systems shows that the recognition loss due to Lombard effect for MOA and vowels reduces compared to the unimodal systems, while the POA recognition is still better due to Lombard effect. A bimodal system is proposed using only alternate acoustic and visual cues which gives a better discrimination of the place and manner of articulation than even standard ASR system. Among the multimodal ASR systems studied, the proposed trimodal system based on Lombard speech gives the best recognition accuracy of 98%, 95%, and 76% for the vowels, MOA and POA, respectively, with an average improvement of 36% over the unimodal ASR systems and 9% improvement over the bimodal ASR systems.

Keywords: Lombard speech; multimodal ASR; throat microphone; visual speech; Convolutional Neural Network; Hidden Markov Model; late fusion; intermediate fusion.

1. Introduction

In adverse circumstances that mostly involve noise, complimentary features such as acoustic speech from a variety of alternate speech sensors, visual speech (lip movements), and gaze effect can help in improving the performance of ASR systems. Hence, combining evidences from these sensors is expected to enhance the performance of ASR systems. The presence of noise affects the performance of an ASR system in two ways: (1) adds environmental noise as an additive component, distorting the speech signal, and (2) induces

Lombard effect in the speaker by altering the speech production mechanism (Lombard, 1911). The speech produced in a noisy environment with more vocal effort is termed as Lombard speech. In this paper, the term "neutral speech" refers to the speech collected in a noise free environment, while Lombard speech refers to the speech collected by feeding crowd noise through the headphone of a speaker to induce Lombard effect into the speech. In order to build robust ASR systems under adverse conditions, it is necessary to address the background additive noise effect as well as the Lombard effect on the performance of these systems. This study attempts to address the exclusive influence of Lombard effect on ASR systems, as explained later.

Various methods to neutralise the effect of additive noise on ASR systems are available in literature. Building robust speech systems for noisy environments using a multimodal approach with alternate speech sensors has been studied extensively. Features from an accelerometer placed at the throat were combined with features from a standard normal microphone (NM) in (Roucos *et al.*, 1986), while features of the throat microphone (TM) and noisy NM speech were combined to estimate clean NM features in (Graciarena *et al.*, 2003). Jou *et al.* (2004) reported ASR recognition of soft whisper from a TM using adaptation methods on a standard speech recogniser. However, in these works effect of Lombard speech on the performance of ASR systems as well as bimodal ASR systems that exclusively use alternate audio-visual speech cues alone were not studied.

The changes induced by the Lombard effect on the acoustic-phonetic characteristics of the normal acoustic microphone speech and visual speech is evident from literature (Rajasekaran *et al.*, 1986; Junqua, Anglade, 1990; Lane, Tranel, 1971; Drugman, Dutoit, 2010; Pisoni *et al.*, 1985; Alexanderson, Beskow, 2014; Davis *et al.*, 2006; Heracleous *et al.*, 2013; Garnier, Henrich, 2014). In noisy conditions, the acoustic-phonetic differences between neutral speech (obtained under noise free laboratory condition) and Lombard speech cause a degradation in the performance of standard ASR systems (that are built under laboratory conditions using neutral speech) due to a mismatch in the training and testing conditions. Methods to compensate the negative impact of Lombard effect on the performance of ASR systems include multistyle training, Lombard speech processing techniques, and feature and model compensation approaches (Boril, 2008; Boril, Hansen, 2010; Bou-Ghazale, Hansen, 1994; Hansen, 1994; Hansen, Varadarajan, 2009; Hansen, Bria, 1990; Sadasivam *et al.*, 2015). Only very few studies have considered bimodal approach for Lombard effect compensation. The bimodal ASR systems studied utilised acoustic and visual speech cues, where the visual Lombard effect is found to degrade the performance of the ASR system in (Heracleous *et al.*, 2013). However, an opposite trend was noticed by (Marxer *et al.*, 2018) with a better claim on accuracy due to a larger audio-visual speech corpus collected from 54 speakers. These studies have used the visual information that remains relatively unaffected by noise, but have not considered the availability of alternate audio information through skin and bone conduction that also remains relatively unaffected by noise. It is necessary to explore the alternate speech related data available in both the visual and audio domains. This would also help to study the feasibility of ASR systems in situations where speech from standard NM is not possible.

The advances in machine learning algorithms as well as increased processing power of computer hardware have led to efficient training algorithms for acoustic modelling of speech using Neural Network (NN) (Hinton *et al.*, 2012). NNs outperform traditional ASR systems based on Hidden Markov Model (HMM), Gaussian Mixture Model (GMM). One such popular and light-weight NN is Convolutional Neural Network (CNN) which contains local, temporal, and spatial filters along the time and frequency domain and showed better results in (Sainath *et al.*, 2013; Abdel-Hamid *et al.*, 2012; Palaz *et al.*, 2013). The spatio-temporal correlations of a signal can be well captured with CNN architecture and it also reduces the translational variance in signals. With fewer parameters, CNN can model the translational invariance, and the speaking style variations and channel distortions are handled with the aid of the maxpooling function.

When compared to phonemes, the CV units have a longer duration, and hence occur with lower frequency in continuous speech. This results in lack of sufficient training examples for these CV units. Hence building robust recognition models for these CV units is an important step towards building robust ASR systems in the context of Indian languages which are syllabic in nature (Shahina, 2007; Khan *et al.*, 2003). To the best knowledge of the authors, there is no study on understanding the Lombard effect on syllable recognition in the Indian languages using a deep learning approach on multimodal ASR systems.

This work aims at studying the exclusive impact of Lombard effect on the recognition of confusable Hindi syllabic (consonant-vowel) units from CNN based unimodal ASR systems based on speech related cues from normal microphone (NM) speech, throat microphone (TM) speech, and image sequences of lip movements to help build robust multimodal ASR systems. This study is further extended to understand the impact of Lombard effect on three bimodal systems and a trimodal system, each for neutral and Lombard speech, built using the three complimentary speech cues. This work also studies the viability of a bimodal system us-

ing only cues from alternate speech sensors (TM + visual speech) in situations where NM may not be available. The percentage accuracy scores obtained from the CNN based unimodal and multimodal ASR systems have been validated using HMM based multimodal ASR systems as well as multihead CNN using intermediate fusion built for this study.

The paper is organised as follows. Section 2 describes the process involved in collecting the acoustic and visual speech data, Sec. 3 explains the features and different unimodal and multimodal ASR systems used in this study. The experimental design and analysis of the results obtained are discussed in Sec. 4. Section 5 summarises the work.

2. The acoustic and visual speech corpus

This section discusses the corpus used for the study.

2.1. Database

The corpus used in this study has a collection of both neutral and Lombard speech samples recorded using both standard NM and TM to record the audio signals, and a camera for capturing corresponding lip movements for the 145 consonant-vowel (CV) units of the Hindi language which are considered as more confusable than even the E-Set vocabulary (Shahina, Yegnanarayana, 2007). The neutral speech and Lombard speech are recorded in different sessions in order to avoid speaker fatigue. In each session, NM and TM speech along with the video are recorded simultaneously. The four kinds of audio and two visual speech signals recorded thus include: NM neutral speech, NM Lombard speech, TM neutral speech, TM Lombard speech, visual neutral speech, and visual Lombard speech. To study the exclusive influence of Lombard effect on the ASR performance, the Lombard speech recordings are carried out by feeding crowd noise (to simulate a crowded, noisy environment) through the headphone of the speaker. The noise played to the speaker through the headphone makes it possible to induce Lombard effect in the recordings while eliminating the additive effect of noise. This makes it possible to record NM, TM, and lip movements that are influenced exclusively by Lombard effect. Such a dataset enables studying exclusively the Lombard effect on the ASR systems. The speech data are collected from seven speakers, three male and four female ones. The corpus contains 145 CV units of the Hindi language. To take into account the intra-speaker and inter-speaker variations in utterances, a total of 20 300 utterances of CV units are used for each of the recording conditions leading to 81 200 utterances from 7 speakers in the complete dataset. 75% of the data are used for training and the remaining 25% are used for testing. The dataset is further grouped into three broad categories, namely, MOA, POA, and vowels, with seven sub-categories of MOA, five sub-categories of POA, and five vowels (Khan *et al.*, 2003). For the Hindi language (and many other Indian languages) the vowel category consists of /*i*/, /*e*/, /*a*/, /*o*/, and /*u*/. The seven manners of articulation (MOA) considered here are UnVoiced UnAspirated (UVUA), UnVoiced Aspirated (UVA), Voiced UnAspirated (VUA), Voiced Aspirated (VA), SemiVowels (SV), Nasals (N), and Fricatives (F). The five different places of articulation (POA) subcategories include velar, palatal, alveolar, dental, and bilabial. Automatic syllable recognition systems are built using these 145 syllables categorised based on the vowels, MOA, and POA.

3. Multimodal ASR system for Hindi syllable recognition

This section discusses the CNN-based unimodal and multimodal ASR systems built for the study. Unimodal, bimodal, and trimodal ASR recognition systems using Convolutional Neural Network (CNN) architecture are built to classify the syllabic CV units based on MOA, POA, and vowels using neutral and Lombard speech cues from standard NM and TM Lombard speech along with their associated visual cues.

A system built in this study is represented by λ^{xyz}, where:

- x represents the sensor – standard normal microphone (N) or throat microphone (T), or camera to capture video signal (V);
- y represents speech data used for training – neutral speech (N) or Lombard speech (L);
- z represents speech data used for testing – neutral speech (N) or Lombard speech (L).

3.1. Unimodal CNN syllable recognition system

Independent CNNs, so called convolutional heads, are used to extract the speech cues from audio and visual streams. To extract the speech cues from each acoustic sensor (normal and throat microphone) two independent one-dimensional CNNs with single channel are utilised, one for normal speech and another one for Lombard speech. Similarly, a two-dimensional CNN is used for visual lip movements extracted from short video clips of CV utterances from the speakers. Each of the six independent unimodal CNN acoustic and visual syllable recognisers to categorise 145 Hindi syllables into three groups, namely MOA, POA, and vowels considered in this study are trained with NM neutral speech, NM Lombard speech, TM neutral speech, TM Lombard speech, visual neutral speech, and visual Lombard speech, respectively. Each system is tested

against their corresponding neutral speech and Lombard speech test data (for matched train-test conditions only). The visual features provide relatively lower performance than acoustic signals even in the matched conditions. Since it is evident from the earlier studies in the literature that unmatched train-test Lombard conditions reduce the performance of speech-input based automatic recognition systems, they are not considered in this study. Table 1 shows all the acoustic and visual unimodal systems with their train, test conditions, and their corresponding symbols used henceforth in this paper.

Table 1. Symbols and notations of unimodal acoustic syllable recognition system using normal and throat microphone.

Sensor	Train data	Test data	Symbol
Normal microphone	Neutral speech	Neutral speech	λ^{NNN}
	Lombard speech	Lombard speech	λ^{NLL}
Throat microphone	Neutral speech	Neutral speech	λ^{TNN}
	Lombard speech	Lombard speech	λ^{TLL}
Camera	Neutral speech	Neutral speech	λ^{VNN}
	Lombard speech	Lombard speech	λ^{VLL}

The unimodal, bimodal, and trimodal audio-visual CNN syllable recognition system using late fusion technique proposed in this study is presented in Fig. 1.

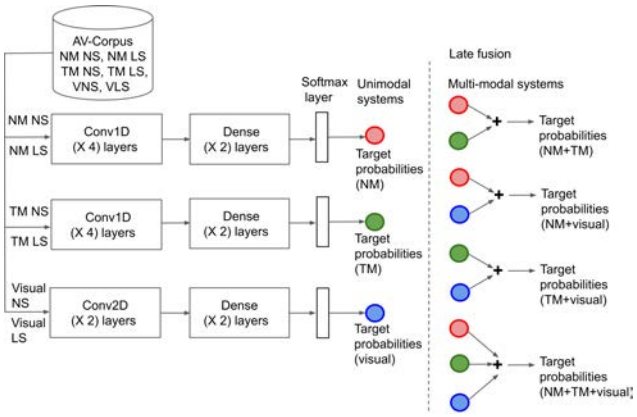


Fig. 1. CNN unimodal, bimodal, and trimodal audio-visual syllable recognition system using late fusion.

The classification layer with two fully connected feed-forward layers makes use of *ReLU* activation function on each hidden unit h_j, in the hidden layer to map the weighted sum of inputs from the previous layer to a scalar output y_j

$$y_j = ReLU(x_j) = \max(0, x_j),$$
$$\text{where} \quad x_j = b_j + \sum_i y_i w_{ij}, \tag{1}$$

where y_i is the output of the unit in the previous layer, w_{ij} is the weight matrix connecting to unit j from unit i in the previous layer. The softmax function in the output layer returns the discrete probability distribution over all classes of MOA/POA/vowels. The class probability output from the softmax layer for each of the unimodal systems λ^{***} is given by

$$p(x_i)_c = \frac{e^{x_i}}{\sum_k e^{x_k}}, \tag{2}$$

where $c \in$ [MOA/POA/vowel], k is an index over all subclasses of MOA/POA/vowel, x is the input vector and $p(x_i)_{\mathrm{MOA}}$, $p(x_i)_{\mathrm{POA}}$, $p(x_i)_{\mathrm{vowel}}$ are the output class probability for unimodal MOA, POA, and vowel systems. There are 7 subclasses of MOA, 5 subclasses of POA, and 5 subclasses of vowels, as mentioned in Sec. 2.

3.2. Multimodal CNN syllable recognition system using late fusion technique

A neutral/Lombard speech bimodal system is designed by combining any two of the three neutral speech/Lombard speech unimodal systems, respectively. For each of the three sound units considered for recognition, two trimodal systems are built by combining the three neutral speech and three Lombard speech unimodal systems. Table 2 shows all the multimodal systems built. Under late fusion, features from different modalities are used to train corresponding unimodal systems and their target probabilities are combined. It can be observed from Fig. 1 that the target probabilities from two independent unimodal CNN systems are combined to perform multimodal syllable recognition.

The bimodal probability score for the systems $\lambda^{(\mathrm{N+T})**}$, $\lambda^{(\mathrm{N+V})**}$, and $\lambda^{(\mathrm{T+V})**}$, are given by

$$p(x_i)_c^{(\mathrm{N+T})**} = p(x_i)_c^{\mathrm{N}**} + p(x_i)_c^{\mathrm{T}**}, \tag{3}$$

Table 2. Symbols and notations of neutral speech and Lombard speech based multimodal syllable recognition systems.

System	Symbol
Bimodal – fusion of (λ^{NNN} and λ^{TNN}) (λ^{NLL} and λ^{TLL})	$\lambda^{(\mathrm{N+T})\mathrm{NN}}$, $\lambda^{(\mathrm{N+T})\mathrm{LL}}$
Bimodal – fusion of (λ^{NNN} and λ^{VNN}), (λ^{NLL} and λ^{VLL})	$\lambda^{(\mathrm{N+V})\mathrm{NN}}$, $\lambda^{(\mathrm{N+V})\mathrm{LL}}$
Bimodal – fusion of (λ^{TNN} and λ^{VNN}), (λ^{TLL} and λ^{VLL})	$\lambda^{(\mathrm{T+V})\mathrm{NN}}$, $\lambda^{(\mathrm{T+V})\mathrm{LL}}$
Trimodal – fusion of (λ^{NNN}, λ^{TNN} and λ^{VNN}), (λ^{NLL}, λ^{TLL} and λ^{VLL})	$\lambda^{(\mathrm{N+T+V})\mathrm{NN}}$, $\lambda^{(\mathrm{N+T+V})\mathrm{LL}}$

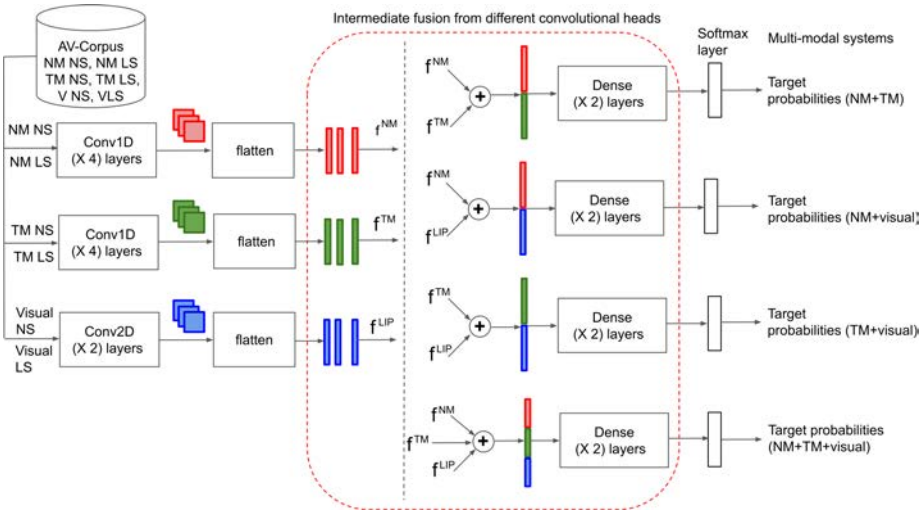


Fig. 2. Multihead CNN based on bimodal and trimodal audio-visual syllable recognition system using intermediate fusion.

$$p(x_i)_c^{(\mathrm{N+V})**} = p(x_i)_c^{\mathrm{N}**} + p(x_i)_c^{\mathrm{V}**}, \quad (4)$$

$$p(x_i)_c^{(\mathrm{T+V})**} = p(x_i)_c^{\mathrm{T}**} + p(x_i)_c^{\mathrm{V}**}, \quad (5)$$

where $** \in [\mathrm{NN/LL}]$, indicating the matched train test conditions of the bimodal systems with neutral and Lombard speech. The trimodal probability score for the systems $\lambda^{(\mathrm{N+T+V})\mathrm{NN}}$, and $\lambda^{(\mathrm{N+T+V})\mathrm{LL}}$, are given by

$$p(x_i)_c^{(\mathrm{N+T+V})\mathrm{NN}} = p(x_i)_c^{\mathrm{NNN}} + p(x_i)_c^{\mathrm{TNN}} + p(x_i)_c^{\mathrm{VNN}}, \quad (6)$$

$$p(x_i)_c^{(\mathrm{N+T+V})\mathrm{LL}} = p(x_i)_c^{\mathrm{NLL}} + p(x_i)_c^{\mathrm{TLL}} + p(x_i)_c^{\mathrm{VLL}}. \quad (7)$$

3.3. Multihead CNN syllable recognition system using intermediate fusion technique

In intermediate fusion the kernel functions representing the feature space of unimodal systems are combined. The bimodal and trimodal audio-visual syllable recognition systems using intermediate fusion technique proposed in this study are presented in Fig. 2. Each feature map from the final pooling layer of each independent CNN head is flattened into a one-dimensional feature vector. The feature spaces of the normal microphone, throat microphone and visual features are represented as $\mathbf{f}^{\mathrm{NM}}$, $\mathbf{f}^{\mathrm{TM}}$, and $\mathbf{f}^{\mathrm{LIP}}$, respectively. These feature vectors from two or more modalities are combined iteratively and fed as input to the following dense layers to find the non-linear mapping function between speech cues from multiple sensors.

4. Experimental setup of unimodal and multimodal recognition of CV units

The recognition accuracy obtained for different Hindi syllabic units under three broad categories, namely, vowel, MOA, and POA, for unimodal, bimodal, and trimodal syllable recognition systems built using CNN is discussed in this section. The results obtained from the CNN based systems are then validated with HMM based systems (late fusion) as well as multihead CNN systems (intermediate fusion).

4.1. Model description of CNN systems

Convolution 1-D, max-pooling, and *ReLU* activation unit are used throughout our experiments for audio stream. The neutral speech is modelled using an architecture containing 4 CNN layers each with a filter size of 2, 5, 20, and 5, respectively. The kernel sizes are 50, 120, 130, and 200 for the each of the CNN layer chronologically. This is followed by two dense layers each of 256 and 128 dimensions, respectively. The speech from the throat microphone is modelled with 4 CNN layers, each with a filter size similar to CNN modelling neutral speech and the kernel sizes are modified to 50, 500, 200, and 100 for each layer, respectively. For modelling video data, two layers of 2-D convolutions are employed, each with 32 and 64 filter size. The kernel sizes are 3×3 for both these layers. The multihead CNN architecture used in this work gathers the output of neutral speech, throat speech, video based CNN model and concatenates it before sending it to the dense layers followed by classification. Categorical cross-entropy loss objective is used in multi-head CNN.

4.2. Model description of HMM systems

To validate the results of the CNN-based systems HMM models are built, one corresponding to each of the CNN-based systems. A 5-state L-to-R HMM, each with 24 Gaussian mixtures, is empirically chosen for building HMM acoustic models, while 12-state L-to-R HMM models with 3 Gaussian mixtures for each state are used for visual speech recognition. All the multimodal systems are designed based on late

fusion technique that combines the log-likelihood scores. The NM and the TM speech signals are represented by 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), along with 13 delta coefficients and 13 delta-delta coefficients, representing the change in spectral content during phonetic transition that could provide cues for phone identity. The 2-dimensional Discrete Wavelet Transform (DWT) coefficients that represent the lip region images corresponding to the sequence of lip movements are used as visual features. DWT is preferred as it captures both the temporal and frequency resolution in a signal. The 39-dimensional audio features and the DWT coefficients are used to build the unimodal ASR systems. For the bimodal and trimodal ASR systems using late fusion technique, the overall conditional log-likelihood, is obtained using the combination of individual log-likelihoods resulting from each unimodal recognition.

4.3. Comparative recognition results of CNN and HMM unimodal systems

The recognition accuracy of vowel, MOA, and POA classes of sounds for the CNN and HMM unimodal systems built with neutral and Lombard speech from NM, TM, and visual speech is given in Fig. 3. In NM speech unimodal systems, for both neutral and Lombard speech, vowels are better recognised compared to the MOA and POA sound categories. However, the TM and visual speech unimodal systems, for both neutral and Lombard speech, give better accuracy for MOA class of sounds than vowel and POA categories. All the unimodal systems, both neutral and Lombard speech, invariably have more poor recognition accuracy for POA category of sound units than that for MOA and vowels. The neutral speech based NM and TM microphone unimodal systems show a marginally higher recognition rate than their Lombard speech based unimodal counterparts for vowel and MOA classes of sounds. However, the POA category of sounds is recognised better in Lombard speech than in neutral speech, refer to Table 3. The visual cues from Lombard speech are recognised better than visual cues from neutral speech for all three (namely, vowel, MOA, and POA) categories of sound, refer to Table 3. This could be attributed to the changes in the vocal effort due to the Lombard effect bringing in more pronounced distinction in the places and manner of articulation in the visual speech, in general. A similar trend is observed in the HMM systems based on Lom-

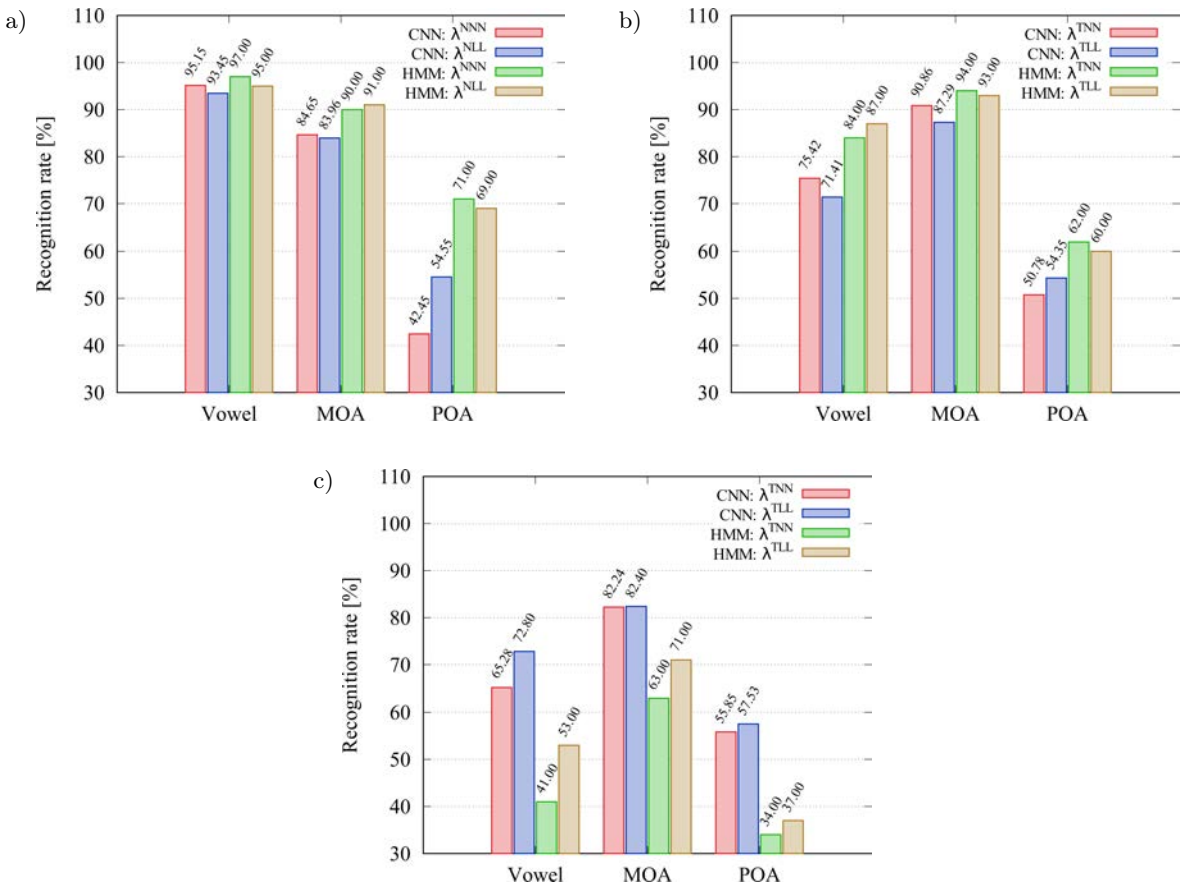


Fig. 3. Comparative recognition of neutral speech *vs* Lombard speech based CNN and HMM unimodal systems built using: a) NM speech – CNN, HMM, b) TM speech – CNN, HMM, c) visual speech – CNN, HMM for matched conditions ($\lambda^{*\mathrm{NN}}$ and $\lambda^{*\mathrm{LL}}$) for vowel, MOA, and POA categories.

Table 3. Percentage of gain (↑) or loss (↓) in accuracy for unimodal visual ASR systems based on Lombard speech. The ↑ or ↓ depict the positive or negative influence of Lombard effect on visual (both CNN and HMM) ASR systems.

Sound units/models	$\lambda_{\text{CNN}}{}^{\text{VLL}}$	$\lambda_{\text{HMM}}{}^{\text{VLL}}$
MOA	no change	8%↑
POA	2%↑	3%↑
Vowel	8%↑	13%↑

bard speech cues in the improvement of recognition accuracy of POA and vowel sound categories as compared to their neutral speech counterparts, refer to Fig. 3. Among the ASR systems based on acoustic cues alone, the vowels are better recognised by the NM based systems, while MOA sound units are better recognised by the TM based systems. This observation is consistent in both the CNN and HMM based systems. The MOA and vowel sound units are, however, relatively worse recognised with visual speech based systems. However, a reverse is observed for the POA category. The visual cues seem to be better discriminating the place of articulation of sound units than both the NM and TM acoustic cues. This trend in unimodal system is observed for both neutral speech and Lombard speech, as seen in Fig. 3. The boost in the performance due to Lombard effect on the CNN-based unimodal visual Lombard speech ASR system over the corresponding HMM-based unimodal visual Lombard speech ASR system is shown in Table 3.

The Lombard effect has a positive influence on the visual cues as seen in Table 3. Both CNN and HMM based systems show the same trend. Though the percentage gain (over the corresponding neutral systems) due to Lombard effect seems to be higher for HMM systems, the performance accuracy of CNN exceeds that of HMM by more than 10% (e.g., CNN gives 82% for MOA, while HMM gives 71%).

Though the recognition trends are similar for both the neutral speech and Lombard speech cues as for which sound category (MOA/POA/vowel) is better captured by which sensor (NM/camera/TM), it is observed that for all sound categories the Lombard effect on both the NM and TM acoustic cues reduces the recognition of MOA and vowels, but improves the recognition of POA category. However, the Lombard effect on the speech related visual cues seems to improve the recognition of the vowels as well as the POA, while MOA remains relatively unaffected. Among the three (two acoustic and one visual) cues, the Lombard effect seems to positively impact the visual cues the most, while also improving the POA recognition in both the acoustic cues. The above discussion shows evidence of complimentary information among the three types of speech in both the neutral and Lombard speech cues. Also, Lombard effect seems to positively impact the recognition of some sound categories. These observations necessitate the study of Lombard effect on ASR systems built using combined evidence.

4.4. Recognition results of multimodal systems

Bimodal systems (NM+TM $\lambda^{(\text{N+T})**}$, NM+video $\lambda^{(\text{N+V})**}$, and TM+video $\lambda^{(\text{T+V})**}$) are built for both neutral and Lombard speech, separately. A trimodal system $\lambda^{(\text{N+T+V})**}$ that combines the two acoustic and the visual cues is also proposed.

4.4.1. Comparative recognition results of CNN multimodal systems using late and intermediate fusion methods

The bimodal and trimodal CNN systems are studied using both late and intermediate fusion techniques. The comparison of recognition accuracy of the bimodal and trimodal CNN systems using late fusion and intermediate fusion techniques for neutral and Lombard speech of MOA, POA, and vowel sound units are given in Table 4. There is only a marginal difference in the recognition accuracy obtained from both the fusion methods for both neutral and Lombard speech based MOA category of sounds, Lombard speech based POA category of sounds and neutral speech based vowel category of sounds for all the multimodal systems. Intermediate fusion technique results in 6.5%, and 3% improvement in recognition accuracy of the neutral speech bimodal CNN systems ($\lambda^{(\text{N+T})\text{NN}}$ and $\lambda^{(\text{V+N})\text{NN}}$) over the corresponding systems built using late fusion method for POA category of sounds. However, for Lombard speech vowel category of sounds, late fusion method shows improvement in recognition accuracy by 2.6%, 3%, and 10% for bimodal CNN systems ($\lambda^{(\text{N+T})\text{NN}}$, $\lambda^{(\text{V+N})\text{NN}}$, and $\lambda^{(\text{V+T})\text{NN}}$) over the

Table 4. Comparison of recognition accuracy of the bimodal and trimodal CNN systems using late fusion and intermediate fusion techniques for neutral and Lombard speech of MOA, POA, and vowel sound units.

System	MOA		POA		Vowels	
	LF	IF	LF	IF	LF	IF
$\lambda^{(\text{N+T})\text{NN}}$	93.68	93.33	54.79	61.36	97.33	97.38
$\lambda^{(\text{N+V})\text{NN}}$	92.67	93.49	61.01	64.02	96.23	95.89
$\lambda^{(\text{T+V})\text{NN}}$	95.28	95.28	68.89	64.68	86.64	87.23
$\lambda^{(\text{N+T+V})\text{NN}}$	96.94	96.72	71.9	74.06	98.51	98.45
$\lambda^{(\text{N+T})\text{LL}}$	92.29	92.84	63.94	64.11	95.39	92.79
$\lambda^{(\text{N+V})\text{LL}}$	92.65	92.04	69.9	69.73	95.27	92.41
$\lambda^{(\text{T+V})\text{LL}}$	93.68	93.75	69.18	68.98	86.1	76.05
$\lambda^{(\text{N+T+V})\text{LL}}$	95.86	95.42	74.31	75.6	96.94	95.62

MOA: manner of articulation, POA: place of articulation, LF: late fusion, IF: intermediate fusion.

corresponding systems built using intermediate fusion method. The trimodal CNN systems for neutral speech POA sound units and Lombard speech vowel sound units show a 2% and 1% improvement in performance under intermediate and late fusion methods, respectively.

From the above experiments denoted in Table 4, we can say that the late fusion technique provides similar gains to those by the intermediate fusion and most of the times better one than that by the intermediate fusion. The late fusion is simple yet effective technique and thus further experiments in this paper make use of the late fusion for system combination.

4.4.2. Comparative recognition results of CNN multimodal systems using late fusion method

The recognition accuracy of vowel, MOA, and POA classes of sounds for the six CNN bimodal systems (three of each for neutral and Lombard speech type) and the two CNN trimodal systems (one of each for neutral and Lombard speech type) for each sound unit using the late fusion method is given in Fig. 4.

The acoustic $\lambda^{(\mathrm{N+T})**}$ bimodal system is the only one among bimodal systems to improve the recognition of all the three sound categories compared to the corresponding individual unimodal ($\lambda^{\mathrm{N}**}$ and $\lambda^{\mathrm{T}**}$) systems (refer Fig. 4). While vowel recognition improves by more than 2% over the NM unimodal system, the MOA and POA recognition improves by 5% and 10% over the TM unimodal system. For the audio-visual ($\lambda^{(\mathrm{N+V})**}$), bimodal systems, the improvement in recognition varies from over 1% for vowels to 8% for MOA and about 20% for POA.

We propose a bimodal system using alternate speech cues ($\lambda^{(\mathrm{T+V})**}$) alone from neutral as well as Lombard speech type. They are studied to understand the feasibility of using Lombard speech based $\lambda^{(\mathrm{T+V})\mathrm{LL}}$ system in adverse environments where NM could be unavailable. The $\lambda^{(\mathrm{T+V})\mathrm{LL}}$ using Lombard speech gives an improvement of over 11% for vowels, 13% for POA,

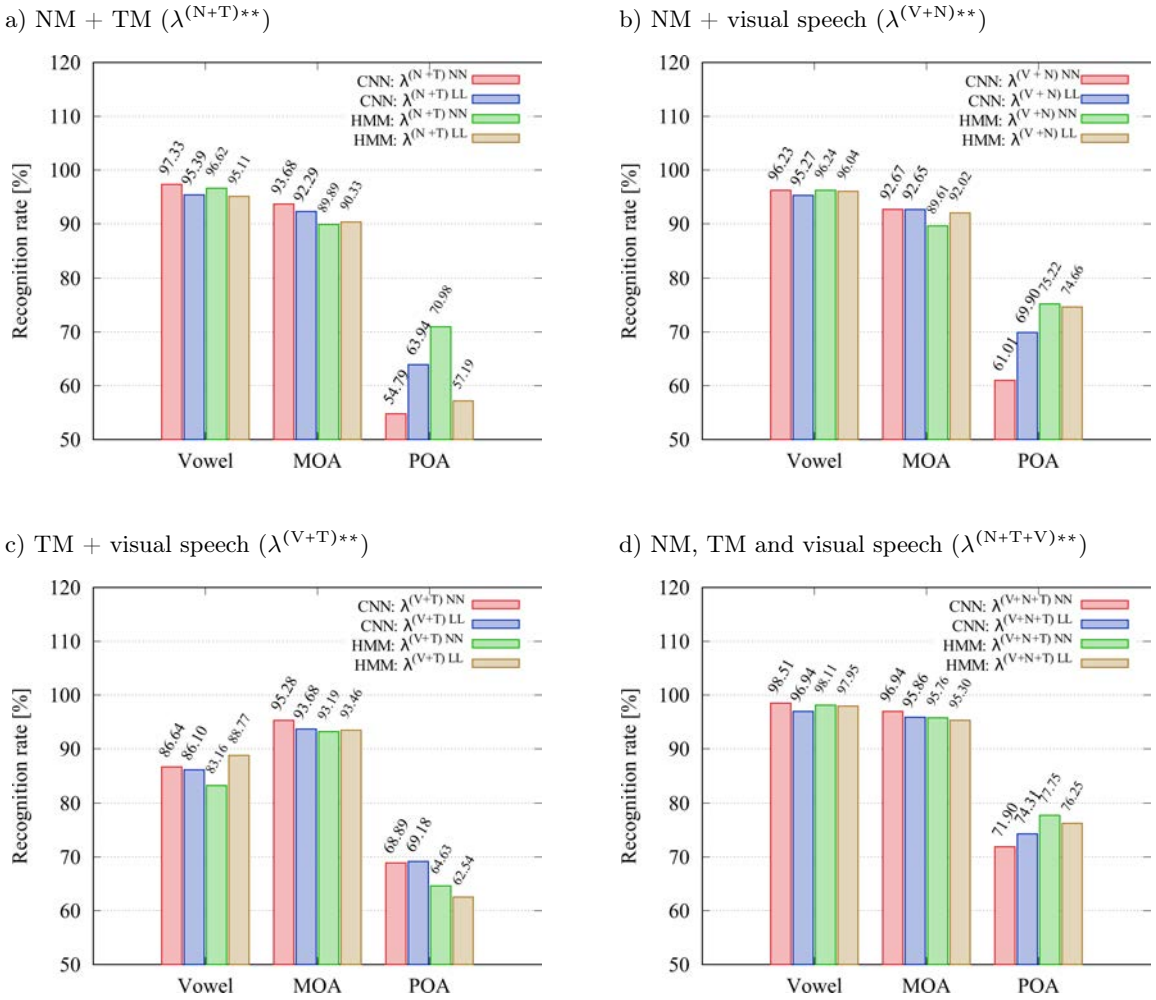


Fig. 4. Comparative recognition of CNN and HMM bimodal systems using: a) standard normal microphone speech and throat microphone speech, b) visual speech and standard normal microphone speech, c) visual speech and throat microphone speech, and d) trimodal systems, for matched conditions ($\lambda^{*\mathrm{NN}}$ and $\lambda^{*\mathrm{LL}}$) of vowel, MOA, and POA categories.

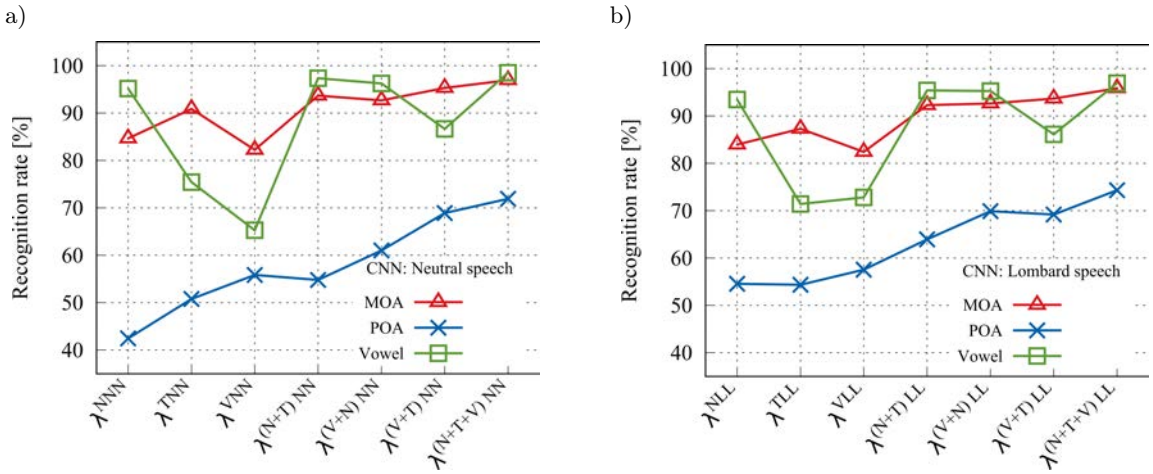


Fig. 5. Comparative recognition of three CNN unimodal systems, three CNN bimodal systems, and a CNN trimodal normal speech (a) and Lombard speech (b) systems for matched conditions ($\lambda^{*\mathrm{NN}}$ and $\lambda^{*\mathrm{LL}}$) of vowel, MOA, and POA categories.

Table 5. Percentage of gain (↑) or loss (↓) in accuracy of the CNN-based Lombard speech ASR systems over the corresponding neutral speech ASR systems.

Sound units/models	λ^{NLL}	λ^{TLL}	λ^{VLL}	$\lambda^{(\mathrm{N+T})\mathrm{LL}}$	$\lambda^{(\mathrm{N+V})\mathrm{LL}}$	$\lambda^{(\mathrm{T+V})\mathrm{LL}}$	$\lambda^{(\mathrm{N+T+V})\mathrm{LL}}$
MOA	<1%↓	3%↓	no change	1%↓	no change	1.5%↓	1%↓
POA	>8%↑	4%↑	2%↑	9%↑	9%↑	< 1%↑	> 3%↑
Vowel	< 2%↓	4%↓	7%↑	2%↓	no change	no change	<2%↓

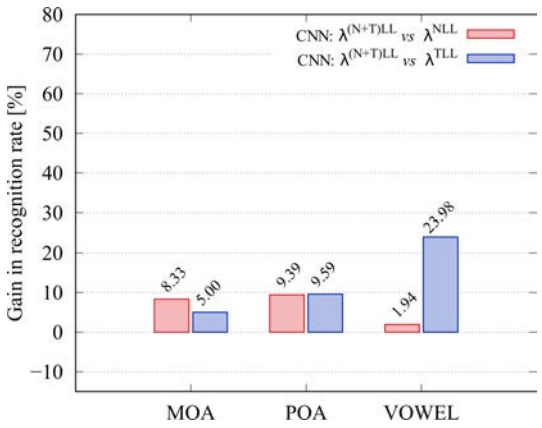
and 5% for vowels over the contributing unimodal systems. This system performs better than even the standard NM systems for both MOA and POA sound categories, with an improvement of over 11% and 15%, respectively, while for vowels alone the performance drops by 7%. Such a trend is observed both for Lombard speech and neutral speech. The drop in recognition of vowels may be overcome with increased training examples and by tuning in the parameters of the model. These results are promising in that ASR systems could be built even in the absence of NM speech in adverse conditions. The proposed trimodal system, $\lambda^{(\mathrm{N+T+V})\mathrm{LL}}$, using Lombard speech gives the best overall performance results over all the previous unimodal and bimodal systems with a comparatively similar performance seen in the neutral speech based trimodal systems for all the three sound categories. All the bimodal and trimodal neutral speech and Lombard speech systems give better results than their individual unimodal systems. This implies that the complimentry speech cues from alternate sensors improve the recognition accuracy of syllable recognition. A comparative recognition accuracy of the vowels, MOA, and POA classes of sounds for the three CNN based unimodal, three CNN based bimodal, and CNN based trimodal systems based on neutral speech along with Lombard speech is given in Fig. 5. The impact of the Lombard effect on the recognition accuracy for all the unimodal, bimodal, and trimodal systems is depicted in Table 5 in terms of percentage gain or loss in accuracy of the CNN-based Lombard speech ASR systems over the corresponding neutral speech ASR systems.

4.4.3. Recognition results obtained from CNN and HMM multimodal systems using late fusion method

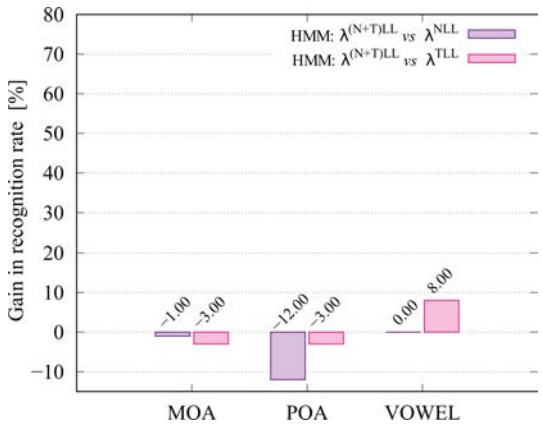
While Fig. 6, shows the percentage of improvement in accuracy of the Lombard speech based bimodal CNN and HMM systems over their unimodal counterparts, Fig. 7 shows the percentage of improvement in accuracy of Lombard speech based trimodal CNN and HMM systems over the unimodal and bimodal systems.

Comparing the Lombard speech CNN-based and HMM-based bimodal systems, for all acoustic systems (NM+TM), even though the HMM model gives hardly any improvement over the standard unimodal NM system, the CNN performs much better for all sound categories with a maximum percentage gain of 15.98% for vowels, see Fig. 6a. For the two audio-visual bimodal systems (NM+visual and TM+visual), both the CNN and HMM systems show significant gain in performance over the standard NM based ASR systems. Though the gain seems higher for HMM systems, the accuracy values are much higher for CNN systems for all the sound categories. The trimodal Lombard speech CNN and HMM based systems also exhibit a similar trend in that they achieve significant gain in performance over the unimodal and bimodal systems.

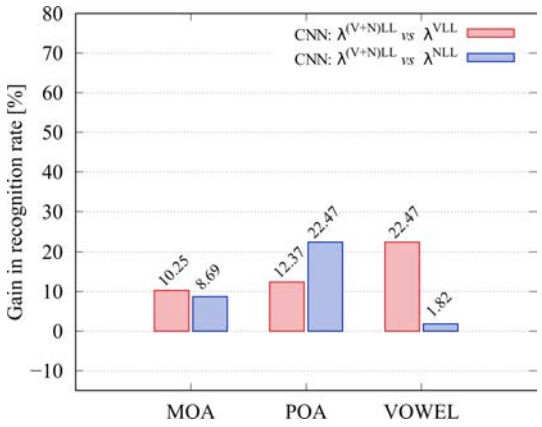
a) CNN – $\lambda^{(N+T)LL}$ *vs* unimodal



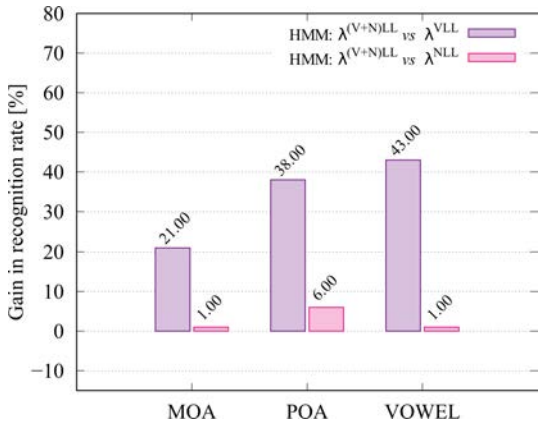
b) HMM – $\lambda^{(N+T)LL}$ *vs* unimodal



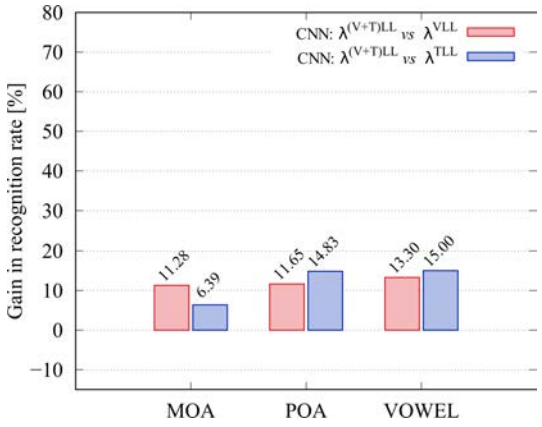
c) CNN – $\lambda^{(V+N)LL}$ *vs* unimodal



d) HMM – $\lambda^{(V+N)LL}$ *vs* unimodal



e) CNN – $\lambda^{(V+T)LL}$ *vs* unimodal



f) HMM – $\lambda^{(V+T)LL}$ *vs* unimodal

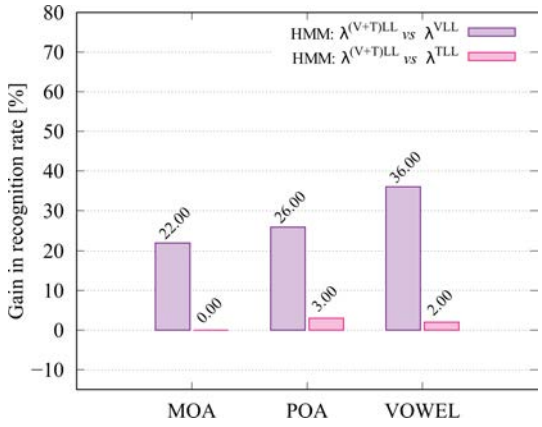
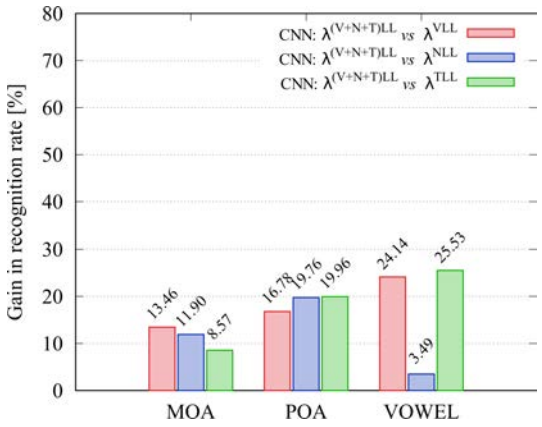


Fig. 6. Percentage of improvement in performance of Lombard speech based bimodal CNN and HMM ASR systems ($\lambda^{(N+T)LL}$, $\lambda^{(V+N)LL}$, and $\lambda^{(V+T)LL}$) over their respective unimodal systems.

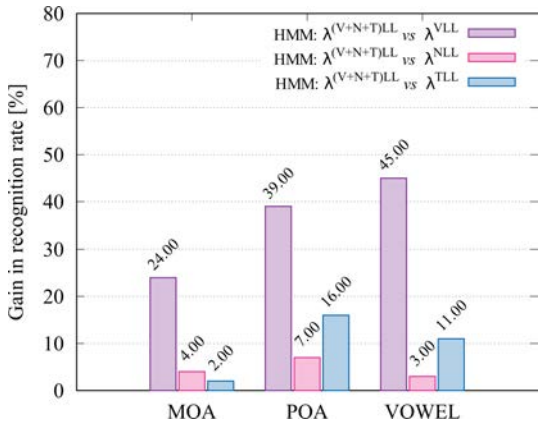
5. Conclusion

This work studies the exclusive influence of Lombard effect on unimodal, bimodal, and trimodal CNN ASR systems built using both standard and alternate sensors. This work also proposes a Lombard speech based bimodal ASR system built using alternate speech cues alone, which gives much better recognition of the place and manner of syllable articulation than the standard ASR system. The dataset built for

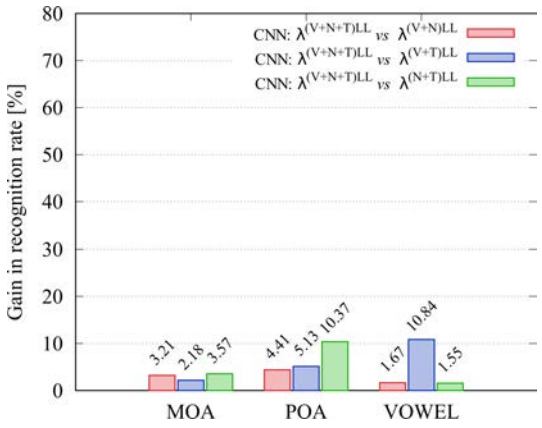
a) CNN – trimodal *vs* unimodal



b) HMM – trimodal *vs* unimodal



c) CNN – trimodal *vs* bimodal



d) HMM – trimodal *vs* bimodal

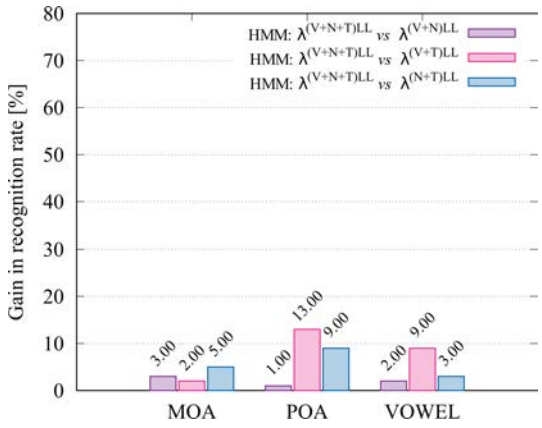


Fig. 7. Percentage of improvement in performance of Lombard speech based trimodal CNN and HMM ASR system ($\lambda^{(V+N+T)LL}$) over the bimodal and unimodal systems.

this study comprising of neutral as well as Lombard speech, recorded from three simultaneous audio and visual cues using normal acoustic microphone, throat microphone, and a camera was designed to study the Lombard effect on the syllable recognition, as Indian languages are syllabic in nature. The unimodal systems built using Lombard speech showed that while Lombard effect helps in better discrimination of place of articulation in both the acoustic cues, it also helps in better discrimination of the manner of articulation and vowels as well in the visual cues. The unimodal systems also showed the different cues containing complimentary information. These speech related cues, when combined to form three different bimodal systems using the late fusion and intermediate fusion approach, gave an improvement in performance over the unimodal systems. The CNN bimodal and trimodal systems implemented using both late fusion and intermediate fusion gave almost similar results. Combining the two acoustic cues alone showed that the Lombard effect boosts the recognition of place of articulation while only marginally degrading the MOA and vowel recognition. When visual cues were combined with either of the audio cues, the Lombard effect had a negligible adverse impact on the recognition of MOA and vowels, while further boosting the place of articulation recognition. A bimodal ASR system based only on alternate speech (TM+visual) cues proposed in this work surprisingly gave better performance of the place and manner of articulation. The trimodal ASR system proposed in this work using the three (two acoustic and one visual) cues gave the best performance, with Lombard effect only marginally affecting the manner of articulation and vowel recognition by less than 2%, while boosting the place of articulation by 3% over the trimodal neutral speech based system. This study was designed to focus on the exclusive influence of Lombard effect on syllable recognition in the context of Indian languages, and also paves the way for building multimodal ASR systems with or without normal microphone under noisy conditions.

References

1. Abdel-Hamid O., Mohamed A., Jiang H., Penn G. (2012), Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, [in:] *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277–4280, doi: 10.1109/ICASSP.2012.6288864.
2. Alexanderson S., Beskow J. (2014), Animated Lombard speech: motion capture, facial animation and visual intelligibility of speech produced in adverse conditions, *Computer Speech & Language*, **28**(2): 607–618, doi: 10.1016/j.csl.2013.02.005.
3. Boril H. (2008), *Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora*, Ph.D. thesis, Czech Technical University in Prague, Czech Rep., https://personal.utdallas.edu/~hynek/.
4. Boril H., Hansen J.H. (2010), Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments, *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(6): 1379–1393, doi: 10.1109/TASL.2009.2034770.
5. Bou-Ghazale S.E., Hansen J.H. (1994), Duration and spectral based stress token generation for hmm speech recognition under stress, [in:] *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94*, Vol. 1, pp. I/413–I/416, doi: 10.1109/ICASSP.1994.389268.
6. Davis C., Kim J., Grauwinkel K., Mixdorff H. (2006), Lombard speech: auditory (A), visual (V) and AV effects, [in:] *Proceedings of the Third International Conference on Speech Prosody*, Citeseer, pp. 248–252.
7. Drugman T., Dutoit T. (2010), Glottal-based analysis of the Lombard effect, [in:] *Interspeech*, pp. 2610–2613.
8. Garnier M., Henrich N. (2014), Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?, *Computer Speech & Language*, **28**(2): 580–597, doi: 10.1016/j.csl.2013.07.005Get.
9. Garnier M., Henrich N., Dubois D. (2010), Influence of sound immersion and communicative interaction on the Lombard effect, *Journal of Speech, Language, and Hearing Research*, **53**(3): 588–608, doi: 10.1044/1092-4388(2009/08-0138).
10. Graciarena M., Franco H., Sonmez K., Bratt H. (2003), Combining standard and throat microphones for robust speech recognition, *IEEE Signal Processing Letters*, **10**(3): 72–74, doi: 10.1109/LSP.2003.808549.
11. Hansen J.H. (1994), Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect, *IEEE Transactions on Speech and Audio Processing*, **2**(4): 598–614, doi: 10.1109/89.326618.
12. Hansen J.H., Bria O.N. (1990), Lombard effect compensation for robust automatic speech recognition in noise, [in:] *First International Conference on Spoken Language Processing*, pp. 1125–1128, https://www.isca-speech.org/archive/icslp_1990/i90_1125.html.
13. Hansen J.H., Varadarajan V. (2009), Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, **17**(2): 366–378, 2009, doi: 10.1109/TASL.2008.2009019.
14. Heracleous P., Ishi C.T., Sato M., Ishiguro H., Hagita N. (2013), Analysis of the visual Lombard effect and automatic recognition experiments, *Computer Speech & Language*, **27**(1): 288–300, doi: 10.1016/j.csl.2012.06.003.
15. Hinton G. *et al.* (2012), Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, **29**(6): 82–97, doi: 10.1109/MSP.2012.2205597.
16. Jou S.-C., Schultz T., Waibel A. (2004), Adaptation for soft whisper recognition using a throat microphone, [in:] *Eighth International Conference on Spoken Language Processing*, pp. 1493–1496, https://www.isca-speech.org/archive/interspeech_2004/i04_1493.html.
17. Junqua J.-C., Anglade Y. (1990), Acoustic and perceptual studies of Lombard speech: application to isolated-words automatic speech recognition, [in:] *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90*, Vol. 2, pp. 841–844, doi: 10.1109/ICASSP.1990.115969.
18. Khan A.N., Gangashetty S.V., Yegnanarayana B. (2003), Syllabic properties of three Indian languages: implications for speech recognition and language identification, [in:] *International Conference on Natural Language Processing*, pp. 125–134.
19. Lane H., Tranel B. (1971), The Lombard sign and the role of hearing in speech, *Journal of Speech, Language, and Hearing Research*, **14**(4): 677–709, doi: 10.1044/jshr.1404.677.
20. Lombard E. (1911), The sign of the elevation of the voice [in French: Le signe de l'élévation de la voix], *Annales des Maladies de l'Oreille, du Larynx, du Nez et du Pharynx*, **37**(2): 101–119.
21. Marxer R., Barker J., Alghamdi N., Maddock S. (2018), The impact of the Lombard effect on audio and visual speech recognition systems, *Speech Communication*, **100**: 58–68, doi: 10.1016/ j.specom.2018.04.006.
22. Palaz D., Collobert R., Magimai-Doss M. (2013), Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks, *CoRR*, Vol. abs/1304.1018, online, http://arxiv.org/abs/1304.1018.
23. Pisoni D., Bernacki R., Nusbaum H., Yuchtman M. (1985), Some acoustic-phonetic correlates of speech produced in noise, [in:] *IEEE International Conference on Acoustics, Speech, and Signal Process-*

ing, ICASSP'85, Vol. 10, pp. 1581–1584, doi: 10.1109/ICASSP.1985.1168217.

24. Rajasekaran P., Doddington G., Picone J. (1986), Recognition of speech under stress and in noise, [in:] *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'86*, Vol. 11, pp. 733–736, doi: 10.1109/ICASSP.1986.1169207.

25. Roucos S., Viswanathan V., Henry C., Schwartz R. (1986), Word recognition using multisensor speech input in high ambient noise, [in:] *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'86*, pp. 737–740, doi: 10.1109/ICASSP.1986.1169208.

26. Sainath T.N., Mohamed A., Kingsbury B., Ramabhadran B. (2013), Deep convolutional neural networks for LVCSR, [in:] *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8614–8618, doi: 10.1109/ICASSP.2013.6639347.

27. Shahina A. (2007), *Processing throat microphone speech*, Ph.D. thesis, Indian Institute of Technology, Madras.

28. Shahina A., Yegnanarayana B. (2007), Mapping speech spectra from throat microphone to close-speaking microphone: a neural network approach, *EURASIP Journal on Advances in Signal Processing*, **2007**: 087219, doi: 10.1155/2007/87219.

29. Sadasivam U.M., Shahina A., Khan A.N., Divya J. (2015), Spectral transformation of Lombard speech to normal speech for speaker recognition systems, [in:] *International Conference Soft Computing Systems*.